ON VALID TESTS OF LINGUISTIC HYPOTHESES

By

D.D. KOSAMBI, Poona

It is known that in any connected piece of writing [“language stream”] the number of words used twice is far less than that used only once. The number occurring three times is still less, and the drop continues rapidly. The Harward philogist George Kingsley Zipf has proposed a “law” for this, the number of words used n times being, according to him, proportional to n2 (1, 24: 2, 40.44). The main purpose of this note is to raise serious objections to this inverse square “law”. These objections are statistical. I maintain that no such law, whatever the exponent, will do for the data so given because the fit is not sufficiently good even when the best exponent is taken by calculations on the logarithmic scale. (1, 25-26; 2, 43; 5, 63). To put this in non-technical language: to every head, There will be the cube-shaped wooden box that fits best, but in general, a rubber cap or a hat of the right size will fit better; and the latter is more likely to indicate a contour of the skull.

1. As my attention was first called to the problem by the Old-Kanarese word counts of Mr. M.G. VENKATESAIYA (working under the direction of Mr. C.R. SANKARAN), I shall illustrate the accepted statistical method by an application to his data. K.V; P denote three works in Halagannada, entitled the Kavirajamarga, Voddaradhane, and Pampasatakam respectively. For purposes of testing it will be necessary to group together the small frequencies at the ends, and sufficient to present the counts as follows.

TABLE 1

Observed				Expected			
Sr	K	V	P	Totals	K	V	P
1	3241	2990	1087	7318	3220-6	3041-3	1056-1
2	270	301	62	633	278-6	263.1	91.3
3	62	71	19	152	66.9	63.2	21.9
4	40	45	14	99	43.6	41.1	14.3
5	29	22	7	58	25.5	24.1	8.4
6	39	47	18	104	45.8	43.2	15.0
	3681	3476	1207	8364	3681-0	3476-0	1207-0

The expected numbers are calculated on the assumption that the three works are uniform in the structure of their language stream, whence it follows that the ratio of the figure in each ‘expected' cell to the total at the foot of its column must be the same as the corresponding ratio of the marginal totals. The numbers obtained are rounded off to the first decimal, taking due care to preserve the totals each

way. As it is clear that the expected and observed totals will never coincide in practice, some method of calculating the magnitude of the discrepancy and of judging its seriousness is necessary. This, for the case in hand, is Karl PEARSON'S X^2 test, X^2 being the sum obtained by squaring the difference between each expectation and observation and dividing the square by the expected number. This sum is here about 22-25, and inasmuch as ten of the given eighteen entries could have been made at will without disturbing the totals, we enter the tables of X^2 (to be found in any standard text on statistics, such as R A. FISTIER'S *Statistical Methods for Research Workers)* with 10 degrees of freedom. It is then found that the probability of exceeding this value X^2 lies between 01 and 02. That is, we should, on the hypothesis of uniformity between the three works, expect to obtain such a result not oftener than once in fifty times but not so rarely as only once in a hundred trials. This is hardly in favour of the hypothesis, though the 'level of significance' is to some extent a matter of individual choice, just as the fit of a hat would depend upon the wearer. If P were smaller than '05, as it is here, the statistician would take the hypothesis as contradicted, following the standard practice of his trade.

This test is surely more exact than anything suggested by ZIPF (5) or his critics (4), judging from the reference material to which I have access here, If the same test be applied *to* the data for the K and the V, it will be found that the two works are compatible, P being not less than about 0.2 which is not at all serious. That is, the *Kavirajamarga and* the Voddaradhane follow about the same frequency laws, but the *Pampasatakam* is decidedly *of* a different nature. The main cause *of* the discrepancy lies in words, of frequency two, of which the V has too many and the P far too few.

2. Applying this X^2 test *to* ZIPF'S data, we reach the following conclusions: Taking together his numbers *for* Chinese and Plautian Latin with ELDRIDGE'S for American newspaper English (1, 23; 2, 26-28), the value of X^2 is enormous and virtually excludes the *very* notion *of* uniformity. Of the three Peiping Chinese and Plautian Latin are closest together, as would be expected from the fact that ELDRIDGE did not count numerals and proper nouns (2, 25). We note in passing that the totals as given by ZIPF need two corrections, that for Chinese being 3342 instead of his 3332, and for Eldridge's English, 6001 in place of 6002. Testing the two languages counted by ZIPF, however, we find X^2 about 40.8, which for 17 degrees of freedom gives a probability of 001, almost exactly, about one chance in a thousand that the two languages follow the same frequency law the discrepancy raising mainly in frequencies 5 and 15.

Finally, the same test applies to any proposed law of frequency, in particular to the inverse square law. For sufficiently extended counts, the expected number of words occurring n times would be given by $6N/(m)^2$ or$60.9N/n^2$, where N is the total number of distinct words counted. The square of each discrepancy is again divided by the expected number; the ratios are added together for the value of x^2. It will be found that of all the six sets of counts cited here, the 'law' applies best to Chinese. It is

again necessary to group together the smaller frequencies at the end (in testing by x^2 the expected frequency should not in any cell fall much below ten) and for 17 degrees of freedom, I obtain a value of x^2=27.17 whereas the value for POS is 27.587. The fit, then, is hardly satisfactory; the best that can be said about the proposed law is that the data for Chinese does not contradict it so decisively as that for the remaining languages.

3. To apply these simple tests, little knowledge of statistical theory, some of pure mathematics, is required. The labour involved is trifling when it is considered that final conclusions are to be drawn from data far more laboriously compiled and that their validity is to be tested. It is surprising, therefore, to note that nowhere in the work of ZIPF, nor in the criticisms of Joes (4) nor the arguments advanced by an able mathematician like STONE (5, 60-61, 63-64) is there any idea of testing goodness of fit or significance. As the U.S.A. are fortunate in possessing many statisticians of eminences, I shall offer a few suggestions here, and leave it to the philologists to work them out if they see fit to do so.

None of the inverse exponent laws fit at all well, though each exponent may be said to characterize the sample from which it was calculated just as the best fitting cubical box would characterize a skull. For KAEDING'S data (2, 23), the three counts given by ZIPF, as well as the three of Kanarese with which I illustrated the X^2 text, a type B series derived from the Poisson distribution or one of Neyman's "contagious" distributions (6) would be found, to fit far better. But the same series would not do for all the samples any better than the same box or hat for all heads; the statistics would be of a descriptive type, lacking the attractive if fictitious Newtonian simplicity of the inverse square law, supplemented by an appeal to SCHRÖDINGER, HEISENBERF, DIRAC (5, 61). Another interesting possibility, if a Poissonian or type B series is found to fit well, would be of estimating the passive vocabulary of the stream, words not used at all, by extrapolation; the "maximum-likelihood" formulae for estimating the words of zero frequency from a supposed Poisson distribution can be worked out very easily, but are not given here inasmuch as the said distribution, which is virtually a random distribution, does not fit.

A far more serious matter is that of properly randomized sampling. ZIPF and his followers wish to characterize an entire language, sometimes all languages by means of their counts. But the total number of words in the respective language streams is always enormous in comparison with the number that can be counted (with obvious exceptions like Anglo-Saxon or Sumerian); therefore every precaution has to be taken to avoid bias. This again, is a matter to which the statisticians have devoted a good deal of time; standard methods of randomization exist which might very well be considered before the work of counting is begun. It is to be noted that ZIPF's scattering.